INFOTECH journal
ISSN : 2460-1861 (Print), 2615-4250 (Online)
Vol. 9 No. 2, pp. 307-318

PERBANDINGAN METODE CART DAN NAÏVE BAYES UNTUK KLASIFIKASI CUSTOMER CHURN

Rahmat Ryan Adhitya[1], Wina Witanti[2], Rezki Yuniarti[3]
[1,2,3]Informatika, Sains dan Informatika, Universitas Jenderal Achmad Yani
Email: [1]rahmatryana18@if.unjani.ac.id, [2]wni@if.unjani.ac.id, [3]rezki@gmail.com

ABSTRACT
Classification is the process of identifying and grouping an object into the same group or category Classification can be used to group a large-sized dataset, and some commonly used classification methods are CART (Classification And Regression Tree) and Naïve Bayes. This study discusses the comparison of CART and Naïve Bayes methods by measuring accuracy, precision, recall, and f1-score values with 3 scenarios of training and testing dataset distribution. Accuracy, precision, recall, and f1-score measurements are performed using a confusion matrix. The scenarios for training and testing dataset division are 70%, 80%, and 90% of the training dataset. From the results of the study, CART has the highest average accuracy and f1-score of 79.616% and 57.636% respectively, while the highest average accuracy and f1-score of Naïve Bayes are 75.104% and 62.004% respectively.

Keywords: Classification, CART, Naïve Bayes, Confusion Matrix.

ABSTRAK
Klasifikasi adalah proses identifikasi dan pengelompokan suatu objek ke dalam kelompok atau kategori yang sama. Klasifikasi dapat digunakan untuk mengelompokkan suatu dataset yang berukuran besar, beberapa contoh metode klasifikasi yang umum digunakan adalah CART (Classification And Regression Tree) dan Naïve Bayes. Penelitian ini membahas mengenai perbandingan metode CART dan Naïve Bayes dengan mengukur nilai akurasi, presisi, recall, dan f1-score dengan 3 skenario pembagian dataset latih dan uji. Pengukuran akurasi, presisi, recall, dan f1-score dilakukan menggunakan confusion matrix. Skenario pembagian dataset latih dan uji adalah 70%, 80%, dan 90% dataset latih. Dari hasil penelitian, CART memiliki rata-rata akurasi dan f1-score tertinggi berturut-turut sebesar 79,616% dan 57,636%, sedangkan rata-rata akurasi dan f1-score tertinggi Naïve Bayes berturut-turut adalah 75,104% dan 62,004%.

Kata Kunci: Klasifikasi, CART, Naïve Bayes, Confusion Matrix.

Riwayat Artikel :

Tanggal diterima : 15-06-2023

Tanggal revisi : 03-07-2023

Tanggal terbit : 04-07-2023

DOI :

https://doi.org/10.31949/infotech.v9i2.5641

1. PENDAHULUAN

1.1. Latar Belakang

Customer churn adalah suatu fenomena dimana seorang pelanggan berhenti menggunakan layanan atau produk dari suatu perusahaan pada periode waktu tertentu dan kemudian berpindah ke perusahaan pesaing. Costumer churn akan menjadi masalah serius jika tidak ditangani secara serius dikarenakan costumer churn dapat mempengaruhi pertumbuhan suatu perusahaan, selain itu fenomena customer churn secara besar-besaran mengindikasikan terdapat kesalahan dalam layanan atau produk yang disediakan (Irmanda et al., 2019).

Banyak perusahaan berusaha mempertahankan pelanggan lama dan menghindari churn dikarenakan cost atau biaya yang dikeluarkan perusahaan untuk memperoleh pelanggan baru jauh lebih tinggi dibandingkan dengan cost yang diperlukan untuk mempertahankan pelanggan lama yang ada (Ahn et al., 2020), berdasarkan fakta tersebut maka tindakan penanganan customer churn yang sesuai dapat menekan pengeluaran suatu perusahaan dan penanganan yang salah dapat membuat pengeluaran suatu perusahaan meningkat sekaligus menghambat pertumbuhan perusahaan (Bagul et al., 2021).

Dalam usaha mempertahankan pelanggan, perusahaan banyak melakukan cara untuk mencegah churn. Salah satunya adalah dengan memanfaatkan teknik data mining untuk melakukan klasifikasi kelompok pelanggan mana yang berpotennsi untuk melakukan churn (Hanifa et al., 2017). Data mining adalah teknik untuk melakukan pengumpulan dan pencarian pola dalam data. Klasifikasi atau prediksi adalah satu dari banyak metode dalam data mining yang digunakan untuk melakukan pengelompokan pelanggan mana yang berpotensi melakukan churn. Metode klasifikasi diperlukan untuk mengolah data yang berukuran besar, semakin besar dimensi data maka semakin tinggi akurasi yang dapat dihasilkan namun semakin lama dan kompleks proses klasifikasi tersebut dilakukan. Terdapat banyak metode klasifikasi dalam data mining dan tiap-tiap metode memiliki keunggulan dan kelemahanya masing-masing.

Metode klasifikasi paling umum adalah metode pohon keputusan dan metode bayesian. CART adalah salah satu metode dari teknik pohon keputusan yang dikembangkan untuk proses analisis klasifikasi data kategorik dan numerik, CART terdiri dari dua metode yaitu metode pohon regresi dan pohon klasifikasi (Prabawati et al., 2019). Naïve Bayes adalah metode klasifikasi berdasarkan teorema Bayesian yang digunakan untuk membuat sebuah prediksi probabilitas label tertentu, Naïve Bayes banyak digunakan karena efisiensi dan kemampuanya menggabungkan probabilitas dari banyak fitur (Riyanto et al., 2021). Pada penelitian ini, CART dan Naïve Bayes dipilih untuk dibandingkan akurasinya dalam melakukan klasifikasi kasus customer churn. Kedua metode tersebut dipilih berdasarkan alasan kemudahan penggunaan, tingkat akurasi yang tinggi, dan kecocokan metode dengan jenis data yang digunakan.

Pada beberapa penelitian sebelumnya (Praningki & Budi, 2018)(Novendri & Andreswari, 2021), metode yang digunakan untuk melakukan validasi dan pengukuran hasil akurasi adalah dengan menggunakan confusion matrix. Confusion matrix digunakan untuk melakukan pengukuran performa suatu metode klasifikasi dengan membandingkan nilai aktual dan nilai prediksi dari hasil metode klasifikasi. Berdasarkan uraian tersebut, penelitian akan berfokus pada perbandingan metode CART dan Naïve Bayes. Kedua metode tersebut dipilih karena tingkat akurasinya yang relatif tinggi. Dalam penelitian ini akan digunakan beberapa rasio pembagian dataset latih dan uji yang telah ditentukan untuk digunakan dalam metode klasifikasi, tujuan dilakukanya hal tersebut adalah untuk menguji pengaruh rasio pembagian dataset latih dan dataset uji terhadap tingkat akurasi metode klasifikasi.

1.2. Tinjuan Pustaka

a. Penelitian Terdahulu

Penelitian terdahulu adalah studi yang dilakukan sebelumnya dan berfungsi sebagai acuan dalam penelitian ini. Dalam acuan tersebut, diambil beberapa teori dan metode dari penelitian sebelumnya. Beberapa referensi yang digunakan dalam penelitian ini menggunakan dataset Telco Customer Churn dan menerapkan metode Naïve Bayes sebagai metode data mining (Kaharudin et al., 2019) (Sjarif et al., 2019) (Halibas et al., 2019) (Hadyan Tisantri et al., 2019) (Yulianti & Saifudin, 2020) (Nalatissifa & Pardede, 2021).

Selain itu, dalam penelitian ini juga digunakan beberapa referensi yang membantu dalam melakukan perbandingan antara metode CART dan Naïve Bayes (Subarkah et al., 2017) (Praningki & Budi, 2018) (Insan et al., 2020). Referensi-referensi tersebut memberikan kontribusi berharga dalam analisis keseluruhan penelitian.

Berdasarkan hasil penelitian-penelitian terdahulu, ditemukan bahwa jumlah dan jenis atribut dalam dataset dapat memengaruhi tingkat akurasi metode data mining yang sedang dibandingkan. Oleh karena itu, penelitian lebih lanjut perlu dilakukan untuk membandingkan metode data mining yang lebih baik dalam melakukan klasifikasi pada jenis data yang berbeda.

b. Customer Churn

Definisi customer churn adalah fenomena dimana pelanggan berhenti menggunakan layanan atau produk dari suatu perusahaan untuk periode waktu tertentu. Pelanggan yang melakukan churn akan berpindah dari satu perusahaan ke pesaing lainnya, hal ini adalah tantangan utama bagi persaingan perusahaan yang kompetitif (Irmanda et al., 2019).

c. Data Mining

Data mining adalah proses penggalian informasi yang berguna dari sebuah kumpulan data yang berukuran besar yang melibatkan analisis data dan penemuan pengetahuan dari database dengan menggunakan pendekatan multi-sisi yang mencangkup analisis statistic, visualisasi data, penemuan pengetahuan, pengenalan pola dan manajemen basis data (Nikmatun & Waspada, 2019).

Data mining dalam proses Knowledge Discovery in Database (KDD) dilakukan dengan langkah-langkah sebagai berikut (Pradana, 2018): (1)Data cleaning, (2)Data integration, (3)Data selection, (4)Data transformation, (5)Data mining, (6)Pattern evaluation, (7)Knowledge representation.

Data mining dapat dilakukan dengan berbagai macam teknik yang berbeda berdasarkan dari hasil mining. Secara umum terdapat tiga jenis teknik yang dapat digunakan dalam data mining untuk mengolah data untuk dipelajari yaitu (Novendri & Andreswari, 2021): (1)Supervised Learning, (2)Unsupervised Learning, (3)Semisupervised Learning.

d. Klasifikasi

Klasifikasi adalah bagian dari teknik supervised learning yang bertujuan untuk melakukan analisa terhadap dataset latih dan mengembangkan sebuah model yang dapat melakukan prediksi untuk memperkirakan kelas yang tidak diketahui dari suatu objek(Tangirala, 2020). Dalam proses klasifikasi data terdapat dua data yang digunakan yaitu: dataset latih (training dataset), atau dataset berlabel yang digunakan saat proses pelatihan untuk membangun model, dan dataset uji (testing dataset) atau dataset tak berlabel yang digunakan untuk memprediksi labelnya dan menentukan tingkat akurasi dari sebuah model. Rasio pembagian dataset latih dan uji secara umum adalah dataset latih lebih besar dari dataset uji dan rasio dataset latih yang sering digunakan bervariasi pada rentang pembagian 70-90% (Alverina et al., 2018)(Praningki & Budi, 2018)(Novendri & Andreswari, 2021).

e. Hyperparameter Tuning

Hyperparameter dalam pembelajaran mesin adalah jenis variabel yang dapat diatur secara langsung sebelum proses pelatihan model oleh pengguna untuk mengatur proses pembelajaran, sementara itu parameter atau parameter model adalah jenis variabel yang didapatkan oleh model saat proses pelatihan (Mantovani et al., 2018)(Yu & Zhu, 2020). Tuning dilakukan untuk mengoptimalkan model dari suatu pembelajaran mesin (Elgeldawi et al., 2021).

Contoh hyperparameter yang digunakan pada CART adalah jumlah sample untuk percabangan (min sample split) dan jumlah daun (min sample leaf) (Mantovani et al., 2018). Contoh hyperparameter Naive Bayes adalah jenis smoothing dan jenis teknik yang digunakan. Contoh parameter pada CART adalah threshold pada setiap cabang, dan contoh parameter pada Naive Bayes adalah nilai likelihood tiap atribut dan prior tiap kelas.

Penelitian ini menggunakan hyperparameter tuning dengan melakukan percobaan menggunakan grid search (Mantovani et al., 2018). Laplace dan Lidstone Smoothing digunakan sebagai hyperparamater tuning untuk Naïve Bayes, jumlah sample percabangan digunakan sebagai hyperparameter tuning untuk CART dengan nilai sample yang digunakan diantara 1 hingga 40 dan jumlah daun yang digunakan diantara 1 hingga 20 (Mantovani et al., 2018).

f. Smoothing

Smoothing adalah teknik untuk menyesuaikan estimasi probabilitas dari suatu model untuk mengatasi data yang tidak ada pada saat pelatihan model, tanpa smoothing probabilitas dapat bernilai 0 jika menemukan data yang tidak muncul saat pelatihan (Yang & Shami, 2020). Probabilitas 0 dapat menyebabkan masalah saat melakukan klasifikasi data. Selain untuk menghindari probabilitas 0, smoothing dapat meningkatkan performa suatu metode (Setyaningsih & Listiowarni, 2021).

Laplace smoothing adalah teknik smoothing yang sering digunakan pada metode Naïve Bayes, nama lain laplace smoothing adalah add-one smoothing dikarenakan tekniknya menambahkan nilai alpha 1 pada setiap token frekuensi (Setyaningsih & Listiowarni, 2021). Teknik smoothing lain adalah lidstone smoothing yang menggunakan prinsip serupa dengan laplace smoothing dengan perbedaan nilai alphanya yang lebih dari 0 dan kurang dari 1, nilai alpha lidstone smoothing yang umum digunaakan adalah 0.1 (Oseki et al., 2019). Perhitungan pada saat proses smoothing dapat dilihat pada Persamaan (1) (Vatanen et al., 2010) (Yang & Shami, 2020).

$$P(X|Y) = \frac{xi+alpha}{x+alpha*k} \quad (1)$$

Dimana xi adalah frekuensi kategori pada atribut x dengan kelas y, x adalah total frekuensi atribut x dengan kelas y, alpha adalah nilai smoothing yang ditentukan, dan k adalah jumlah kelas atau label.

g. Binning

Binning adalah salah satu metode dalam transformasi data untuk mengelompokkan data kedalam bin atau kriteria tertentu. Metode ini digunakan pada tahap persiapan data untuk metode-metode klasifikasi seperti Naïve Bayes. Metode binning dilakukan dengan melakukan pemeriksaan terhadap nilai-nilai yang ada pada sekelilingnya (Nguyen & Zucker, 2019). Salah satu metode binning adalah equal-width binning yang melakukan partisi berdasarkan jarak antar bin, untuk melakukan binning berdasarkan jarak antar

bin maka nilai maksimum dan minimum suatu data harus diketahui, untuk mengetahui nilai maksimum dan minimum dapat dilakukan dengan mengurutkan data dari nilai terkecil ke terbesar. Setelah nilai maksimum dan minimum diketahui maka jarak dapat diketahui dengan menggunakan Persamaan (2).

$$Jarak = max - min \quad (2)$$

Jarak adalah rentang data terkecil ke terbesar, max adalah nilai terbesar dalam dataset, dan min adalah nilai terkecil dalam dataset. Untuk mengetahui interval antar bin maka dapat dilakukan dengan menggunakan Persamaan (3).

$$Interval = \frac{Jarak}{n\ bin} \quad (3)$$

Interval adalah jarak antar bin dan n bin adalah jumlah bin. Sedangkan untuk mengetahui batas interval dapat dilakukan dengan menggunakan Persamaan (4).

$$Batas\ Interval = min + (k-1) * Interval \quad (4)$$

Batas Interval adalah batas atas atau bawah untuk tiap bin dan k adalah urutan bin. Selain digunakan untuk mengelompokkan data, metode binning juga digunakan untuk meminimalisasi kesalahan dan meningkatkan akurasi dari model prediksi.

h. Information Gain (IG)

IG adalah salah satu metode yang digunakan untuk melakukan pemilihan atribut atau fitur dengan mengidentifikasi atribut mana yang paling relevan (Al-Harbi, 2019), Pemilihan atribut dilakukan dengan melakukan ranking atribut berdasarkan nilai IG terbesar ke terkecil dan mengesampingkan fitur yang tidak memenuhi batas nilai IG yang telah ditentukan (Hasibuan & Marji, 2019).

Nilai IG didapatkan dari perhitungan total entropy kriteria pada suatu atribut dikurangi dengan entropy masing-masing kriteria. Entropy merepresentasikan tingkat heterogenitas dalam dataset, semakin tinggi nilai entropy maka semakin tinggi tingkat heterogenitas dalam dataset. Perhitungan entropy dapat dilihat pada Persamaan (5).

$$Entropy(S) = -\sum_{i=1}^{n} p_i \log_2 p_i \quad (5)$$

S adalah himpunan kasus, n adalah jumlah partisi S, dan pi adalah proporsi Si terhadap S. Setelah perhitungan entropy dilakukan, maka proses perhitungan IG dapat dilakukan dengan rumus yang ditunjukan pada Persamaan (6).

$$Gain(S,A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entropy(S_i) \quad (6)$$

S adalah himpunan kasus, A adalah atribut, n adalah jumlah partisi atribut A, |Si| adalah proporsi Si terhadap S, |S| adalah jumlah kasus dalam S, Entropy(S) adalah entropy sebelum pemisah (atribut), dan Entropy(Si) adalah entropy setelah pemisah (kriteria).

Setelah IG tiap atribut telah dihitung maka dapat dilakukan rangking dari atribut dengan nilai IG terbesar ke terkecil, lalu nilai batas atau threshold harus ditentukan untuk melihat atribut mana yang memenuhi syarat untuk digunakan. Menurut beberapa penelitian, terdapat beberapa metode untuk menentukan nilai threshold yang diantaranya adalah: (1)Menentukan nilai threshold secara independent atau menggunakan nilai yang lebih umum seperti 0,05 (Prasetiyowati et al., 2021), (2)Menentukan n% dari atribut yang akan digunakan (25%, 50%, 75% dari atribut) atau n atribut tertinggi (10 atribut dengan nilai IG tertinggi) (Bolón-Canedo & Alonso-Betanzos, 2019).

i. Gini Index (GI)

GI adalah suatu pengukuran yang menunjukkan seberapa sering suatu elemen yang dipilih secara acak dari sekumpulan elemen akan diberi label yang salah jika diberi label secara acak sesuai dengan distribusi label di subset(Ghasemi et al., 2020). GI juga digunakan sebagai salah satu metode yang digunakan untuk melakukan pemilihan atribut atau fitur dengan mengukut tingkat kemurnian suatu fitur terhadap kelas yang ada(Al-Harbi, 2019). Perhitungan GI dapat dilihat pada Persamaan (7).

$$GI(S) = 1 - \sum_{i=1}^{n} pi^2 \quad (7)$$

N adalah jumlah partisi dan pi adalah proporsi Si terhadap S, jika atribut A pada sebuah data terbagi menjadi dua kategori A1 dan A2 maka perhitungan GI dapat dilihat pada Persamaan (8).

$$GI(A) = \frac{N1}{N} GI(A1) + \frac{N2}{N} GI(A2) \quad (8)$$

N1 adalah jumlah atribut A bernilai A1, N2 adalah jumlah atribut A bernilai A2, dan N adalah jumlah atribut A.

j. Classification And Regression Tree (CART)

CART adalah singkatan dari Classification And Regression Trees (pohon klasifikasi dan regresi). CART diperkenalkan oleh Breiman pada tahun 1984. Metode CART dapat membangun pohon klasifikasi maupun pohon regresi. Pohon klasifikasi dibangun oleh CART dengan membagi atribut secara biner. Index Gini digunakan untuk memilih atribut yang akan dibagi. Metode CART juga digunakan untuk analisis regresi menggunakan pohon regresi. Fitur regresi dari CART dapat digunakan untuk melakukan forecasting variabel dependen yang memiliki kumpulan variabel prediktor pada periode waktu tertentu (Arora et al., 2017).

Kelebihan metode CART: CART dapat mengatasi missing value secara otomatis menggunakan surrogate splits, CART dapat menggunakan kombinasi variabel diskrit dan kontinu, CART secara otomatis melakukan seleksi variabel, CART dapat membentuk interaksi antar variabel.

Kekurangan metode CART: Pohon keputusan CART mungkin tidak stabil, CART hanya dibagi oleh satu variable, Non-parameter.

Adapun langkah-langkah metode CART adalah (Jones & Makmun, 2021): (1)Menyusun calon cabang (candidate split). Penyusunan dilakukan pada semua variabel prediktor. Daftar yang berisi calon cabang disebut calon cabang mutakhir. (2)Menghitung nilai Gini calon cabang kanan dan kiri, lalu menghitung Gini total untuk calon cabang tersebut. (3)Menentukan calon cabang mana yang akan dijadikan cabang dengan memilih calon cabang dengan nilai Gini terkecil. Jika tidak terdapat simpul keputusan, proses CART akan dihentikan. Namun, jika terdapat simpul keputusan, dilanjutkan kembali ke langkah kedua, dengan terlebih dahulu membuang calon cabang yang sudah berhasil menjadi cabang sehingga mendapatkan calon cabang mutakhir terbaru.

Perhitungan nilai Gini untuk calon cabang kanan dan kiri dapat menggunakan Persamaan (8) dan setelahnya untuk menghitung total Gini calon cabang tersebut dapat menggunakan Persamaan (9) (Praningki & Budi, 2018)(Insan et al., 2020).

k. Naïve Bayes

Naïve Bayes adalah metode klasifikasi statistik yang menghitung kemungkinan kesamaan antara kasus lama dan baru berdasarkan kasus, Naïve Bayes mampu memprediksi kelas keanggotaan probabilistik dan memiliki akurasi tinggi dan kecepatan yang baik jika diterapkan pada database besar. Naïve Bayes adalah metode supervised learning yang berarti metode ini membutuhkan data awal berupa dataset latih selama fase pembelajaran. Teori dasar Naïve Bayes adalah menghitung probabilitas-probabilitas dari kategori dalam data untuk mendapatkan perkiraan probabilitas dari kategori yang diberikan, persamaan umum yang digunakan untuk teorema naïve bayes adalah seperti pada Persamaan (9) (Praningki & Budi, 2018):

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \quad (9)$$

X adalah kriteria suatu kasus berdasarkan masukan, Y adalah kelas solusi pola label, P(Y|X) adalah probabilitas kemunculan label kelas Y dengan kriteria masukan X, P(X|Y) adalah probabilitas kemunculan label kelas X dengan kriteria masukan Y, P(Y) adalah probabilitas label kelas Y, dan P(X) adalah probabilitas label kelas X.

l. Cross Validation

Cross validation adalah teknik yang digunakan untuk melakukan evaluasi performa model dengan membagi data yang ada kedalam set latihan dan validasi lalu melatih model pada set latihan dan mengevaluasinya dengan set validasi. Teknik ini digunakan untuk mengestimasi kemampuan model untuk menggeneralisasi data yang belum pernah muncul sebelumnya(Widaningsih, 2019).

K-fold cross validation adalah bentuk umum dari cross validation, data dibagi menjadi k partisi dan dilakukan iterasi pada k sebagai set validasi dan k-1 atau sisanya sebagai set latih (Utami et al., 2020). Nilai k beragam dimana semakin besar nilai k maka semakin baik hasil evaluasinya tetapi akan meningkatkan cost atau waktu perhitungan, nilai k yang umum digunakan adalah k=5 dan k=10 (Subarkah et al., 2017)(Widaningsih, 2019)(Hasnain et al., 2020).

m. Confusion Matrix

Confusion matrix adalah matriks untuk perhitungan hasil klasifikasi. Confusion matrix berisi hasil klasifikasi yang dapat digunakan untuk mengukur performa model klasifikasi. Melalui confusion matrix, akurasi, presisi, recall, dan f1-score dapat diketahui. Metrik pengukuran dalam confusion matrix dapat dibagi menjadi dua nilai, yaitu akurasi dan tingkat error. Dengan mengetahui berapa banyak data yang terklasifikasikan dengan benar maka dapat diketahui akurasi hasil prediksi, dan dengan mengetahui berapa banyak data yang terklasifikasikan dengan salah maka dapat diketahui tingkat error dari hasil prediksi (Santra & Christy, 2012).

Pada pengukuran performa model klasifikasi menggunakan confusion matrix, terdapat empat nilai yang digunakan untuk merepresentasikan hasil klasifikasi. Keempat nilai tersebut ialah True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN). True Positive adalah jumlah data positif yang terdeteksi dengan benar, Nilai True Negative adalah jumlah data negatif yang terdeteksi dengan benar, False Positive adalah jumlah data negatif yang terdeteksi sebagai data positif, dan False Negative adalah jumlah data positif yang terdeteksi sebagai data negatif.

Berdasarkan nilai True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN) dapat diperoleh nilai akurasi, error, ketepatan dan nilai penarikan. Akurasi menggambarkan seberapa akurat metode dapat mengklasifikasikan data secara benar. Dengan kata lain, nilai akurasi adalah perbandingan antara data yang terklasifikasikan benar dengan keseluruhan data. Nilai akurasi dapat diperoleh dengan Persamaan (10) (Hasnain et al., 2020).

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} \quad (10)$$

Presisi adalah rasio prediksi benar positif dibandingkan dengan keseluruhan hasil yang diprediksi positif, nilai presisi diperoleh dengan Persamaan (11) (Hary Candana et al., 2021).

$$Precision = \frac{TP}{TP+FP} * 100\% \quad (11)$$

Recall adalah rasio prediksi benar positif dibandingkan dengan keseluruhan data yang benar positif, nilai presisi diperoleh dengan Persamaan (12) (Hary Candana et al., 2021).

$$Recall = \frac{TP}{TP+FN} * 100\% \quad (12)$$

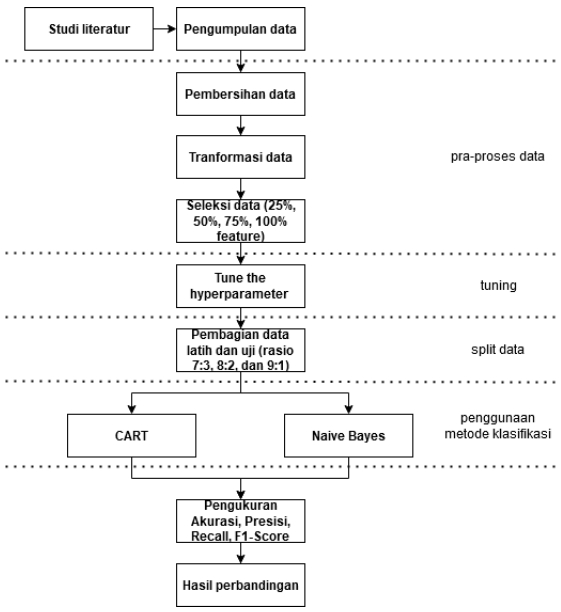
TP(True Positive) adalah jumlah data positif yang terklasifikasi benar untuk kelas ke-i, TN(True Negative) adalah jumlah data negatif yang terklasifikasi benar untuk kelas ke-i, FP(False Positive) adalah jumlah data negatif namun terklasifikasi dengan salah untuk kelas ke-i, dan FN(False negative) adalah jumlah data positif namun terklasifikasi dengan salah untuk kelas ke-i.

F1-score adalah metrik pengukuran yang mengkombinasikan presisi dan recall untuk menghasilkan pengukuran tunggal yang merepresentasikan keefektifan suatu klasifikasi, f1-score adalah rata-rata dari presisi dan recall (Bolón-Canedo & Alonso-Betanzos, 2019). f1-score menyediakan hasil pengukuran dengan presisi dan recall yang seimbang, perhitungan f1-score dilakukan menggunakan Persamaan (13) (Hanifa et al., 2017)(Novendri & Andreswari, 2021)(Halibas et al., 2019)(Bolón-Canedo & Alonso-Betanzos, 2019).

$$F1\ score = \frac{2*precision*recall}{precision+recall} \quad (13)$$

1.3. Metodologi Penelitian

Metode penelitian berisi langkah-langkah yang dilakukan dalam mengklasifikasikan customer churn menggunakan metode CART dan Naïve Bayes serta membandingkan tingkat akurasinya. Adapun tahapan dari metode penelitian ini adalah: studi literature, pengumpulan data, pra-proses data, tuning hyperparameter, pelatihan model, pengujian model, dan analisis hasil seperti yang ada pada Gambar 1.



Gambar 1. Metodologi Penelitian

a. Pengumpulan Data

Pada tahap ini dilakukan pengumpulan data dan informasi mengenai customer churn dan didapatkan sebuah dataset publik yang diambil dari situs Kaggle. Data tersebut adalah data telco customer churn yang memiliki 21 atribut termasuk kelas data atau label (churn dan not churn) dan memiliki 7043 baris data dan terakhir diperbaharui tahun 2018. Data tersebut digunakan untuk melakukan klasifikasi apakah pelanggan berpotensi melakukan churn atau tidak, atau untuk melakukan prediksi perilaku pelanggan dalam upaya mengembangkan program untuk mempertahankan pelanggan.

Dataset telco customer churn terdiri atas 21 atribut termasuk atribut kelas atau label yang direpresentasikan dengan CustomerID, Gender, SeniorCitizen, Partner, Dependents, Tenure, PhoneService, MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, Contract, PaperlessBilling, PaymentMethod, MonthlyCharges, TotalCharges, and Churn.

b. Pra-Proses Data

Pra-proses data adalah tahapan dimana data akan diproses untuk mengurangi noise dan kesalahan data sebelum proses selanjutnya. Atribut CustomerID dan Churn tidak digunakan dalam proses ini karena CustomerID adalah atribut ID unik, sedangkan Churn adalah label. Adapun tahapan proses pra-proses adalah:

a). Data Cleaning / Pembersihan data

Pada tahap ini, data secara manual dibersihkan dari noise, missing value, dan atribut yang tidak relevan menggunakan Microsoft Excel. Setelah dilakukan pemeriksaan, ditemukan bahwa terdapat 11 missing value dan 0 data tidak relevan, semua missing value terdapat di kolom TotalCharges. Karena jumlah missing value kecil (0,001% dari total data atau 11/7043), keputusan diambil untuk menghapus data dengan missing value.

b). Data Transformation / Transformasi data

Data diubah menjadi format yang sesuai menggunakan metode binning untuk mengkonversi atribut numerik menjadi bentuk nominal (Nguyen & Zucker, 2019). Pada tahap ini, tiga atribut, Tenure, MonthlyCharges, dan TotalCharges, yang masih dalam bentuk numerik, diubah menjadi data kategorikal menggunakan metode binning.

c). Data Selection / Seleksi data

Data yang relevan dipilih menggunakan information gain untuk meminimalkan ukuran data dan meningkatkan akurasi hasil data mining (Hasibuan & Marji, 2019). Hasil perhitungan gain informasi dapat dilihat pada Tabel 1.

Tabel 1. Feature Sorting by IG Weight

No	Atribut	Information Gain
1	Contract	0,141647
2	Tenure	0,0933796
3	OnlineSecurity	0,0930946

4	TechSupport	0,0907063
5	InternetService	0,0799162
6	OnlineBackup	0,0673147
7	PaymentMethod	0,0640887
8	DeviceProtection	0,0631674
9	StreamingMovies	0,0460485
10	StreamingTV	0,0458814
11	MonthlyCharges	0,0365071
12	PaperlessBilling	0,0275833
13	TotalCharges	0,0246508
14	Dependents	0,0205877
15	Partner	0,0164224
16	SeniorCitizen	0,0151954
17	MultipleLines	0,0011518
18	PhoneService	0,0000998
19	Gender	0,0000527

Tabel 1 mengurutkan atribut berdasarkan information gain, dengan Contract memiliki nilai tertinggi dan Gender memiliki nilai terendah. Untuk mengurangi dimensi data, atribut dipilih berdasarkan threshold yang ditetapkan pada 25%, 50%, 75%, dan 100% dari nilai information gain tertinggi.

c. Tuning Hyperparameter

Tuning adalah proses menguji konfigurasi hyperparameter yang ditentukan sebelum proses pelatihan klasifikasi, hyperparamater yang akan diuji akan didata dan ditentukan rentang yang akan diuji. Cross validation digunakan untuk mengevaluasi konfigurasi tuning terhadap performa model, konfigurasi tuning dengan performa terbaik digunakan pada saat klasifikasi.

d. Pembagian Data

Setelah tahapan pra-proses dan tuning maka data sudah siap digunakan dan dapat digunakan. Data dibagi menjadi dua, pertama adalah dataset latih untuk pelatihan model, dan dataset uji untuk menguji performa metode CART dan Naïve Bayes. Pembagian rasio data ini dibagi menjadi tiga skenario pembagian yaitu dataset latih 70%, 80%, dan 90%.

e. Penggunaan Metode Klasifikasi

Algoritma pada penelitian ini adalah CART dan Naive Bayes. Model-model dibangun menggunakan dataset latih berlabel yang telah disiapkan. Konfigurasi model telah ditentukan setelah penyetelan hyperparameter. Model yang dihasilkan oleh CART berupa pohon keputusan yang dapat diinterpretasikan sebagai aturan, dan model yang dihasilkan oleh Naive Bayes berupa tabel prioritas dan likelihood untuk setiap atribut. Proses pengujian model dilakukan menggunakan dataset uji yang tidak berlabel untuk menghasilkan hasil prediksi.

f. Pengukuran Akurasi

Pada tahap ini hasil prediksi dan data aktual diproses dalam confusion matrix untuk menghasilkan metrik akurasi, error, presisi, recall, dan f1-score. Confusion Matrix digunakan untuk menghitung akurasi dari hasil suatu klasifikasi. Dataset penelitian ini memiliki dua buah label atau kelas, yaitu kelas churn dan not churn. Maka matriks yang akan terbentuk kurang lebih akan menjadi seperti pada Tabel 2.

Tabel 2. Confusion Matrix

		Kelas Aktual	
		Churn (1)	**!Churn (0)**
Kelas Prediksi	**Churn (1)**	TP	FP
	!Churn (0)	FN	TN

2. PEMBAHASAN

2.1. Hyperparameter Tuning

Pada tahap ini dilakukan tuning terhadap hyperparameter di kedua metode, nilai hyperparameter yang diujikan bersumber pada penelitian-penelitian terdahulu dapat dilihat pada Tabel 3.

Tabel 3. Tuning Parameter

Metode	**Hyperparameter**	**Tuning Value**
Naïve Bayes	Alpha / Smoothing Alpha	0 (no smoothing), 0,1-0,9 (lidstone smoothing), 1 (laplace smoothing)
CART	Min_sample_split	1-40
	Min_sample_leaf	1-20

Pengujian konfigurasi tuning dilakukan menggunakan fungsi GridSearchCV dan 10-fold cross validation, sehingga hasil pencarian tuning value terbaik dapat dilihat pada Gambar 2 untuk CART dan 3 untuk Naïve Bayes.

```
Best Parameters:  {'min_samples_leaf': 20, 'min_samples_split': 1}
Best Accuracy: 0.793
```

Gambar 2. Best Tuning Configuration - CART

```
Best Parameters:  {'alpha': 0}
Best Accuracy: 0.729
```

Gambar 3. Best Tuning Configuration – Naïve Bayes

Gambar 2 menunjukkan CART dengan min_sample_leaf =20 dan min_samples_split =1 adalah hasil tuning terbaik, dan Gambar 3 menunjukkan Naïve Bayes dengan nilai alpha = 0 adalah hasil tuning terbaik, sehingga nilai hyperparameter tersebut digunakan dalam pengujian model selanjutnya.

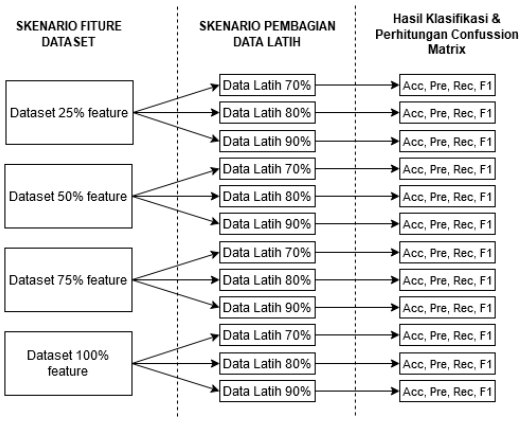
2.2. Skenario Pengujian

Skenario pengujian yang dilakukan pada penelitian ini dapat dilihat pada Tabel 4, dimana dataset yang digunakan berjumlah 4 dimana masing-masing dataset memiliki jumlah fitur yang berbeda sesuai dengan threshold dan urutan information gain yang ditentukan.

Tabel 4. Dataset and Split Ratio Scenario

Dataset	Rasio Split Dataset latih
25% feature (5)	70%, 80%, 90%
50% feature (10)	70%, 80%, 90%
75% feature (14)	70%, 80%, 90%
100% feature (19)	70%, 80%, 90%

Berdasarkan Tabel 4, pengujian model menghasilkan 12 pengukuran akurasi, presisi, recall, dan f1. Untuk gambaran pengujian scenario dapat dilihat pada Gambar 4.



Gambar 4. Dataset and Split Ratio Scenario

Untuk detail masing-masing scenario fiture dataset dapat dilihat pada Tabel 5.

Tabel 5. Dataset Scenario Detail

Dataset	Detail Fiture
25% fiture	Tenure, InternetService, OnlineSecurity, TechSupport, Contract
50% fiture	Tenure, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, Contract, PaymentMethod
75% fiture	Dependents, Tenure, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, Contract, PaperlessBilling, PaymentMethod, MonthlyCharges, TotalCharges
100% fiture	Gender, SeniorCitizen, Partner, Dependents, Tenure, PhoneService, MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, Contract, PaperlessBilling, PaymentMethod, MonthlyCharges, TotalCharges

Perhitungan akurasi metode CART dan Naïve Bayes dilakukan menggunakan confusion matrix yang persamaanya dapat dilihat pada Persamaan (7) hingga (9). Perangkat lunak yang dibuat dapat menampilkan confusion matrix dan perhitungan untuk akurasi, error, presisi, recall dan f1-score.

2.3. Hasil Pengujian dan Perbandingan

a. Akurasi

Akurasi adalah ukuran rasio hasil prediksi yang benar (Churn & Tidak Churn) dengan keseluruhan data.

Nilai akurasi juga dapat diinterpretasikan sebagai sebarapa banyak pelanggan yang benar diprediksi churn dan tidak churn dari keseluruhan pelanggan.

Tabel 6 menunjukkan akurasi rata-rata tertinggi Naïve Bayes berada pada rasio dataset latih 70% pada semua skenario. Akurasi rata-rata tertinggi Naive Bayes adalah 75,104% pada rasio dataset latih 70% skenario 1, sedangkan akurasi rata-rata CART adalah 79,616% pada rasio dataset latih 80% skenario 2.

Selisih akurasi pembagian dataset latih tiap skenario dataset dan rasio metode Naïve Bayes adalah 2,538% dan CART adalah 0,795%, sedangkan selisih akurasi tertinggi antara metode Naïve Bayes dan CART adalah 4,511%.

Tabel 6. Accuracy Comparison

Skenario Dataset	CART			Naïve Bayes		
	Train 70%	Train 80%	Train 90%	Train 70%	Train 80%	Train 90%
1	78,848341	**78,87704**	78,821022	**75,10427**	75,095949	75,056818
2	79,350711	**79,61621**	79,204546	**72,77725**	72,565744	72,642046
3	**79,57346**	79,495379	79,19034	**73,10427**	72,850035	72,755682
4	79,213271	79,402985	**79,40341**	**73,07109**	72,828713	72,755682

b. Presisi

Presisi adalah ukuran rasio prediksi yang benar (positif) dengan seluruh hasil prediksi positif.

Nilai presisi juga dapat diinterpretasikan sebagai dari sekian banyak pelanggan yang diprediksi churn, seberapa banyak yang benar-benar melakukan churn. Nilai presisi menggambarkan seberapa handal model memprediksi kelas positif / churn dengan tetap meminimalisir kesalahan memprediksi kelas tidak churn menjadi churn (False Positive). Semakin tinggi nilai presisi maka semakin akurat model memprediksi kelas positif / churn.

Tabel 7 menunjukkan nilai presisi rata-rata model CART lebih tinggi dari model CART pada semua skenario, sedangkan presisi rata-rata tertinggi Naïve Bayes berada pada rasio 70% pada semua skenario. Presisi rata-rata tertinggi Naive Bayes adalah 52,212% pada rasio dataset latih 70% skenario 1,

sedangkan presisi rata-rata tertinggi CART adalah 66,854% pada rasio dataset latih 80% skenario 1.

Selisih presisi pembagian dataset latih tiap skenario dataset dan rasio metode Naïve Bayes adalah 3,187% dan CART adalah 3,270%, sedangkan selisih presisi tertinggi antara metode Naïve Bayes dan CART adalah 14,642%.

Tabel 7. Precision Comparison

Skenario Dataset	CART			Naïve Bayes		
	Train 70%	Train 80%	Train 90%	Train 70%	Train 80%	Train 90%
1	66,465951	**66,85479**	65,618658	**52,21278**	52,167783	52,121174
2	64,40267	**65,13026**	64,152231	**49,29014**	49,025026	49,149708
3	**64,80426**	64,415714	63,672544	**49,66467**	49,340895	49,261194
4	**64,41964**	64,210661	63,584399	**49,62816**	49,3145	49,252504

c. Recall

Recall adalah ukuran rasio prediksi benar (positif) dengan keseluruhan data yang benar positif.

Nilai recall juga dapat diinterpretasikan sebagai dari sekian banyak pelanggan yang melakukan churn, seberapa banyak yang benar diprediksi churn. Nilai recall menggambarkan sebarapa handal model dalam mengidentifikasi kelas positif / churn, semakin tinggi nilai recall maka semakin tinggi kemampuan model untuk mengidentifikasi kelas positif / churn pada data.

Tabel 8 menunjukkan nilai recall rata-rata model Naïve Bayes lebih tinggi dibandingkan nilai recall rata-rata CART pada semua skenario dan rasio dataset latih, nilai recall rata-rata tertinggi Naïve Bayes.berada pada rasio 90% pada semua skenario. Recall rata-rata tertinggi Naive Bayes adalah 81,497% pada rasio dataset latih 90% skenario 2, sedangkan presisi rata-rata tertinggi CART adalah 52,780% pada rasio dataset latih 90% skenario 4.

Selisih recall pembagian dataset latih tiap skenario dataset dan rasio metode Naïve Bayes adalah 5,561% dan CART adalah 11,229%, sedangkan selisih recall tertinggi antara metode Naïve Bayes dan CART adalah 28,716%.

Tabel 8. Recall Comparison

Skenario Dataset	CART			Naïve Bayes		
	Train 70%	Train 80%	Train 90%	Train 70%	Train 80%	Train 90%
1	42,192511	41,550801	**43,42246**	76,078432	75,935829	**76,57754**
2	50,089127	**50,40107**	49,358288	81,105169	80,882352	**81,49733**
3	50,784313	**51,14973**	50,909092	80,60606	80,160426	**80,69519**
4	49,126559	50,962568	**52,78075**	79,982175	79,705882	**80,05348**

d. F1-Score

F1-Score adalah ukuran gabungan antara presisi dan recall.

Nilai f1-score yang tinggi mengindikasikan keseimbangan antara presisi dan recall yang berarti hasil klasifikasi dapat melakukan prediksi kelas positif dengan akurat dan memiliki cangkupan tinggi pada semua kelas positif aktual.

Tabel 9 menunjukkan nilai f1-score Naïve Bayes lebih tinggi dari CART pada semua skenario dataset dan rasio pembagian dataset latih. F1-score rata-rata tertinggi Naive Bayes adalah 62,004% pada rasio dataset latih 90% skenario 1, sedangkan presisi rata-rata tertinggi CART adalah 57,636% pada rasio dataset latih 90% skenario 4.

Selisih f1-score pembagian dataset latih tiap skenario dataset metode Naïve Bayes adalah 1,079% dan CART adalah 6,557%, sedangkan selisih f1-score tertinggi antara metode Naïve Bayes dan CART adalah 4,367%.

Tabel 9. F1-Score Comparison

Skenario Dataset	CART			Naïve Bayes		
	Train 70%	Train 80%	Train 90%	Train 70%	Train 80%	Train 90%
1	51,3514	51,078908	**51,9676**	61,91055	61,835814	**62,00443**
2	56,283308	**56,77532**	55,740057	**61,30856**	61,044533	61,304086
3	56,925144	**56,97458**	56,488443	**61,45164**	61,079349	61,159775
4	55,70194	56,796774	**57,63669**	**61,23937**	60,925323	60,965573

2.4. Analisa Hasil Pengujian dan Perbandingan

Berdasarkan hasil perhitungan dapat disimpulkan bahwa model CART menghasilkan nilai akurasi dan presisi lebih baik dibanding Naïve Bayes, dimana model CART dengan rasio dataset latih 80% memiliki nilai akurasi rata-rata sebesar 79,616% dan nilai presisi rata-rata sebesar 66,854%, sedangkan akurasi rata-rata tertinggi untuk model Naïve Bayes adalah 75,104% dan nilai presisi rata-rata sebesar 52,212% dengan signifikasi perbedaan akurasi sebesar 4,511% dan presisi sebesar 14,64%.

Performa f1-score Naïve Bayes mengungguli performa f1-score CART, dengan f1-score tertinggi Naïve Bayes adalah 62,004% dan CART adalah 57,636%, kedua nilai f1-score tersebut berada pada rasio dataset latih 90% dengan signifikasi perbedaan f1-score sebesar 4,367%.

Nilai presisi dan recall menjadi penting untuk memprediksi secara akurat dan tepat seberapa banyak pelanggan yang berpotensi churn dan mengurangi hasil prediksi potensi churn ketika seharusnya pelanggan tidak berpotensi churn, dengan adanya nilai f1-score memungkinkan untuk menyeimbangkan nilai recall dan presisi dan mendapatkan nilai terbaik dari keduanya.

3. KESIMPULAN

Berdasarkan hasil uji coba terhadap metode Naive Bayes dan CART, kesimpulan pertama yang dapat ditarik adalah metode CART mengungguli metode Naive Bayes berdasarkan akurasi dan presisi tiap percobaan dan hasil dari akurasi dan presisi rata-rata dari sepuluh percobaan. CART berhasil mengungguli Naïve Bayes berdasarkan hasil rata-rata akurasi pada rasio 70%, 80% dan 90% dataset latih.

Kesimpulan kedua adalalah metode Naïve Bayes mengungguli metode CART berdasarkan recall dan f1-score tiap percobaan dan hasil dari recall dan f1-score rata-rata dari sepuluh percobaan. Naïve Bayes berhasil mengungguli CART berdasarkan hasil rata-rata recall dan f1-score pada rasio 70%, 80% dan 90% dataset latih.

Nilai presisi dan recall menjadi penting untuk memprediksi secara akurat dan tepat seberapa banyak pelanggan yang berpotensi churn dan mengurangi hasil prediksi potensi churn ketika seharusnya pelanggan tidak berpotensi churn, dengan adanya nilai f1-score memungkinkan untuk menyeimbangkan nilai recall dan presisi dan mendapatkan nilai terbaik dari keduanya.

PUSTAKA

Ahn, J., Hwang, J., Kim, D., Choi, H., & Kang, S. (2020). A Survey on Churn Analysis in Various Business Domains. IEEE Access, 8, 220816–220839. https://doi.org/10.1109/ACCESS.2020.3042657

Al-Harbi, O. (2019). A Comparative Study of Feature Selection Methods for Dialectal Arabic Sentiment Classification Using Support Vector Machine. International Journal of Computer Science and Network Security, 19(1), 167–176. https://doi.org/10.48550/arXiv.1902.06242

Alverina, D., Chrismanto, A. R., & Santosa, R. G. (2018). Perbandingan Algoritma C4.5 dan CART dalam Memprediksi Kategori Indeks Prestasi Mahasiswa. Jurnal Teknologi Dan Sistem Komputer, 6(2), 76–83. https://doi.org/10.14710/jtsiskom.6.2.2018.76-83

Arora, A., Gupta, B., Uttarakhand, P., & Rawat, I. A. (2017). Analysis of Various Decision Tree Algorithms for Classification in Data Mining. International Journal of Computer Applications, 163(8), 15–19.

Bagul, N., Berad, P., Surana, P., & Khachane, C. (2021). Retail Customer Churn Analysis using RFM Model and K-Means Clustering. International Journal of Engineering Research & Technology, 10(03), 349–354. https://doi.org/DOI : 10.17577/IJERTV10IS030170

Bolón-Canedo, V., & Alonso-Betanzos, A. (2019). Ensembles for feature selection: A review and future trends. Information Fusion, 52(1), 1–12. https://doi.org/10.1016/j.inffus.2018.11.008

Elgeldawi, E., Sayed, A., Galal, A. R., & Zaki, A. M. (2021). Hyperparameter tuning for machine learning algorithms used for arabic sentiment analysis. Informatics, 8(4), 1–21. https://doi.org/10.3390/informatics8040079

Ghasemi, F., Neysiani, B. S., & Nematbakhsh, N. (2020). Feature selection in pre-diagnosis heart coronary artery disease detection. 6th International Conference on Web Research (ICWR), 6, 27–32. https://doi.org/10.1109/ICWR49608.2020.9122285

Hadyan Tisantri, D., Cahya Wihandika, R., & Adinugroho, S. (2019). Prediksi Keputusan Pelanggan Menggunakan Extreme Learning Machine Pada Data Telco Customer Churn. Jurnal Pengembangan Teknologi Informasi Dan Ilmu KomputerJurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer, 3(11), 10516–10523.

Halibas, A. S., Cherian Matthew, A., Pillai, I. G., Harold Reazol, J., Delvo, E. G., & Bonachita Reazol, L. (2019). Determining the intervening effects of exploratory data analysis and feature engineering in telecoms customer churn modelling. 2019 4th MEC International Conference on Big Data and Smart City, 1–7. https://doi.org/10.1109/ICBDSC.2019.8645578

Hanifa, T. T., Adiwijaya, & Al-faraby, S. (2017). Analisis Churn Prediction pada Data Pelanggan PT. Telekomunikasi dengan Logistic Regression dan Underbagging. E-Proceeding of Engineering, 4(2), 78.

Hary Candana, E. W., Gede, I., Gunadi, A., & Divayana, D. G. H. (2021). Perbandingan Fuzzy Tsukamoto, Mamdini Dan Sugeno Dalam Penentuan Hari Baik Pernikahan Berdasarkan Wariga Menggunakan Confusion Matrix. Jurnal Ilmu Komputer Indonesia, 6(2), 14–22.

Hasibuan, M. R., & Marji. (2019). Pemilihan Fitur dengan Information Gain untuk Klasifikasi Penyakit Gagal Ginjal menggunakan Metode Modified K-Nearest Neighbor (MKNN). Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer, 3(11), 10435–10443. http://j-ptiik.ub.ac.id

Hasnain, M., Pasha, M. F., Ghani, I., Imran, M., Alzahrani, M. Y., & Budiarto, R. (2020). Evaluating Trust Prediction and Confusion Matrix Measures for Web Services Ranking. IEEE Access, 8, 90847–90861.

https://doi.org/10.1109/ACCESS.2020.2994222

Insan, N., Hadijati, M., & Irwansyah, I. (2020). Perbandingan Metode Classification and Regression Trees (CART) dengan Naïve Bayes Classification (NBC) dalam Klasifikasi Status Gizi Balita di Kelurahan Pagesangan Barat. Eigen Mathematics Journal, 3(1), 14. https://doi.org/10.29303/emj.v1i2.68

Irmanda, H. N., Astriratma, R., & Afrizal, S. (2019). Perbandingan Metode Jaringan Syaraf Tiruan Dan Pohon Keputusan Untuk Prediksi Churn. JSI: Jurnal Sistem Informasi (E-Journal), 11(2), 1817–1825. https://doi.org/10.36706/jsi.v11i2.9286

Jones, A. H. S., & Makmun, M. S. (2021). Implementasi Metode CART untuk Klasifikasi Diagnosis Penyakit Hepatitis Pada Anak. Journal of Informatics, Information System, Software Engineering and Applications, 3(2), 61–70. https://doi.org/10.20895/INISTA.V3I2

Kaharudin, Pradana, M. G., & Kusrini. (2019). Prediksi Customer Churn Perusahaan Telekomunikasi Menggunakan Naïve Bayes Dan K-Nearest Neighbor. Jurnal Informasi Interaktif, 4(3), 165–171.

Mantovani, R. G., Horváth, T., Cerri, R., Junior, S. B., Vanschoren, J., & de Carvalho, A. C. P. de L. F. (2018). An empirical study on hyperparameter tuning of decision trees. https://doi.org/https://doi.org/10.48550/arXiv.1812.02207

Nalatissifa, H., & Pardede, H. F. (2021). Customer Decision Prediction Using Deep Neural Network on Telco Customer Churn Data. Jurnal Elektronika Dan Telekomunikasi, 21(2), 122–127. https://doi.org/10.14203/jet.v21.122-127

Nguyen, T. H., & Zucker, J. D. (2019). Enhancing metagenome-based disease prediction by unsupervised binning approaches. Proceedings of 2019 11th International Conference on Knowledge and Systems Engineering, KSE 2019, 1–5. https://doi.org/10.1109/KSE.2019.8919295

Nikmatun, I. A., & Waspada, I. (2019). Implementasi Data Mining untuk Klasifikasi Masa Studi Mahasiswa Menggunakan Algoritma K-Nearest Neighbor. Jurnal SIMETRIS, 10(2), 421–432.

Novendri, R., & Andreswari, R. (2021). Implementasi Data Mining Untuk Memprediksi Customer Churn Menggunakan Algoritma Naive Bayes. E-Proceeding of Engineering, 8(2), 2762–2773.

Oseki, Y., Yang, C., & Marantz, A. (2019). Modeling Hierarchical Syntactic Structures in Morphological Processing. Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, 43–52. https://doi.org/10.18653/v1/w19-2905

Prabawati, N. I., Widodo, & Duskarnaen, M. F. (2019). Kinerja Algoritma Classification a nd Regression Tree (Cart) da lam Mengklasifikasikan Lama Masa Studi Mahasiswa y ang Mengikuti Organisasi d i Universitas Negeri Jakarta Avalaiable at : Avalaiable at : Jurnal Pinter, 3(2), 139–145.

Pradana, E. (2018). Analisis Penerapan Adaptive Boosting (Adaboost) Dalam Meningkatkan Performasi Algoritma C4.5. Jurnal Teknologi Pelita Bangsa, 96.

Praningki, T., & Budi, I. (2018). Sistem Prediksi Penyakit Kanker Serviks Menggunakan CART, Naive Bayes, dan k-NN. Creative Information Technology Journal, 4(2), 83. https://doi.org/10.24076/citec.2017v4i2.100

Prasetiyowati, M. I., Maulidevi, N. U., & Surendro, K. (2021). Determining threshold value on information gain feature selection to increase speed and prediction accuracy of random forest. Journal of Big Data, 8(1), 22. https://doi.org/10.1186/s40537-021-00472-4

Riyanto, E. A., Juninisvianty, T., Nasution, D. F., & Risnandar, R. (2021). Analisis Kinerja Algoritma CART dan Naive Bayes Berbasis Particle Swarm Optimization (PSO) untuk Klasifikasi Kelayakan Kredit Koperasi. Jurnal Teknologi Informasi Dan Ilmu Komputer, 8(1), 55. https://doi.org/10.25126/jtiik.0812988

Santra, A. K., & Christy, C. J. (2012). Genetic Algorithm and Confusion Matrix for Document Clustering. International Journal of Computer Science, 3(2), 322–328. http://ijcsi.org/papers/IJCSI-9-1-2-322-328.pdf

Setyaningsih, E. R., & Listiowarni, I. (2021). Categorization of Exam Questions based on Bloom Taxonomy using Naïve Bayes and Laplace Smoothing. 3rd 2021 East Indonesia Conference on Computer and Information Technology, EIConCIT 2021, 330–333. https://doi.org/10.1109/EIConCIT50028.2021.9431862

Sjarif, N. N. A., Yusof, M. R. M., Wong, D. H. Ten, Ya'akob, S., Ibrahim, R., & Osman, M. Z. (2019). A Customer Churn Prediction using Pearson Correlation Function and K Nearest Neighbor Algorithm for Telecommunication Industry. International Journal of Advances in Soft Computing and Its Applications, 11(2), 46–59.

Subarkah, P., Santiko, I., & Tri, A. (2017). Perbandingan Kinerja Algoritma Cart dan Naive Bayesian untuk Mendiagnosa Penyakit Diabetes Melitus. Conference on Information Technology, Information System and Electrical Engineering, 17.

Tangirala, S. (2020). Evaluating the impact of GINI index and information gain on classification using decision tree classifier algorithm. International Journal of Advanced Computer Science and Applications, 11(2), 612–619. https://doi.org/10.14569/ijacsa.2020.0110277

Utami, Y. T., Shofiana, D. A., & Heningtyas, Y. (2020). Penerapan Algoritma C4.5 Untuk Prediksi Churn Rate Pengguna Jasa Telekomunikasi. Jurnal Komputasi, 8(2), 69–76. https://doi.org/10.23960/komputasi.v8i2.2647

Vatanen, T., Väyrynen, J. J., & Virpioja, S. (2010). Language identification of short text segments with n-gram models. Proceedings of the 7th International Conference on Language Resources and Evaluation, LREC 2010, 3423–3430.

Widaningsih, S. (2019). Perbandingan Metode Data Mining Untuk Prediksi Nilai dan Waktu Kelulusan Mahasiswa Prodi Teknik Informatika Dengan Algoritma C4.5, Naïve Bayes, KNN, dan SVM. Jurnal Tekno Insentif, 13(1), 16–25. https://doi.org/10.36787/jti.v13i1.78

Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. Neurocomputing, 415, 295–316. https://doi.org/10.1016/j.neucom.2020.07.061

Yu, T., & Zhu, H. (2020). Hyper-Parameter Optimization: A Review of Algorithms and Applications. 1–56. https://doi.org/https://doi.org/10.48550/arXiv.2003.05689

Yulianti, Y., & Saifudin, A. (2020). Sequential Feature Selection in Customer Churn Prediction Based on Naive Bayes. IOP Conference Series: Materials Science and Engineering, 879(1), 7. https://doi.org/10.1088/1757-899X/879/1/012090